Iterative Back-translation は対訳語彙を獲得できるか?

谷川 琢磨　　秋葉 友良　　塚田 元
豊橋技術科学大学
{tanigawa.takuma.fu, akiba.tomoyoshi.tk, tsukada.hajime.hl}@tut.jp

概要

ニューラル機械翻訳におけるデータ拡張手法として，Iterative Back-translation (IBT) が知られている．IBT は，翻訳対象言語対の 2 つの単言語コーパスを相互に逆翻訳とモデルの更新を繰り返し行うことで，疑似対訳データと翻訳モデルの質を向上させる手法である．IBT は効果的な手法であることが知られているが，その知識獲得の過程は十分に解明されていない．本研究では，IBT を用いたドメイン適応を対象に，単言語コーパスからどのような過程でターゲットドメインの対訳語彙を獲得しているかについて調査を行った．その結果，反復を繰り返すごとに対訳語彙を獲得していき，最終的には獲得可能な 6 割以上の対訳語彙が獲得できていることが示された．

1　はじめに

近年のニューラル機械翻訳（NMT）における有効的なデータ拡張手法として，Iterative Back-translation(IBT)[1][2] が知られている．IBT は，翻訳対象言語対の 2 つの単言語コーパスを相互に逆翻訳とモデルの更新を繰り返し行うことで，疑似対訳データと翻訳モデルの質を向上させる手法である．

IBT は効果的なデータ拡張手法であることが実験的に示されている一方，その原理については十分に明らかになっているとは言えない．特に，互いに対応関係のない 2 言語の単言語コーパスから，実際に翻訳に関する知識が獲得されるかどうか，詳細に調べた研究は存在しない．そこで本研究では，IBT を用いたドメイン適応の問題設定を対象に，単言語コーパスからドメイン固有の対訳語彙が獲得されるかどうかを調査することで，IBT の知識獲得の過程を明らかにする．その際，対訳語彙を 2 つの単言語コーパスから獲得するには，少なくとも対訳となる単語対それぞれが各単言語コーパスに存在する必要があると考えられるため，理想的な設定として単語対が必ず含まれることが保証されるコンパラブルコーパスを用いた実験も行った．加えて，翻訳モデルの処理単位が対訳語彙の獲得に与える影響を調査するために，サブワード単位および単語単位の翻訳モデルで比較を行った．

2　Iterative Back-translation（IBT）

Iterative Back-translation(IBT) に基づくドメイン適応手法の手順を説明する．ここで，X と Y はそれぞれの言語を示し，言語 X から Y の翻訳を X-Y，Y から X への翻訳を Y-X と記す．

1. ソースドメインの対訳コーパス C_X^{out} と C_Y^{out} を用いて，$Model_{X-Y}0$ と $Model_{Y-X}0$ を学習する．
2. i を 0 に初期化して以下を反復する．
   - 2.1 ターゲットドメインの単言語コーパス C_Y^{in} を $Model_{Y-X}i$ により翻訳し，疑似対訳コーパス（$C_X'^{in}, C_Y^{in}$）を作成する．疑似対訳コーパスと（C_X^{out}, C_Y^{out}）を結合した学習データを用いて，$Model_{X-Y}i$ から Fine-tuning を行い $Model_{X-Y}(i+1)$ を学習する．
   - 2.2 ターゲットドメインの単言語コーパス C_X^{in} を $Model_{X-Y}i$ により翻訳し，疑似対訳コーパス（$C_Y'^{in}, C_X^{in}$）を作成する．疑似対訳コーパスと（C_Y^{out}, C_X^{out}）を結合した学習データを用いて，$Model_{Y-X}i$ から Fine-tuning を行い $Model_{Y-X}(i+1)$ を学習する．
   - 2.3 $i \leftarrow i+1$

3　調査手法

ターゲットドメインの知識獲得の過程を確認するために，本実験では対訳語彙の獲得を調査した．言語 X と言語 Y の対訳語彙獲得を調査するために行った手順は以下のとおりである．

1. 言語 X のソースドメインの学習データ (対訳コーパス) に存在する単語集合 O_X とターゲットドメインの学習データ (単言語コーパス) に存

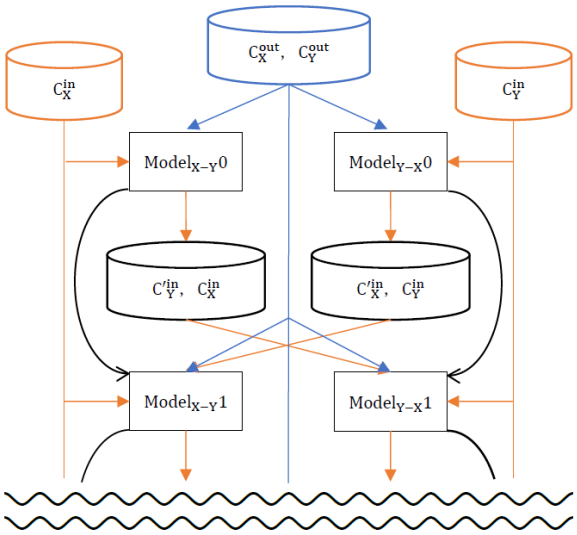

図 1　Iterative Back-translation の手順

在する単語集合 I_X をそれぞれ求め，ターゲットドメインの学習データにのみ存在する単語集合 $D_X = I_X - O_X$ を特定する．同様に言語 Y の $D_Y = I_Y - O_Y$ も求める．

2. Moses[3] の単語アライメントツールを用いて，テストデータにおける言語 X と言語 Y の単語アライメントを求める[1]．テストデータ中でアライメントされた単語対 (w_X, w_Y) について，両言語の単語がともに D_X，D_Y に含まれているもの $T = \{(w_X, w_Y) | w_X \in D_X \wedge w_Y \in D_Y\}$ を獲得目標の対訳語彙とする．
3. 調査対象の翻訳モデルによってテストデータの翻訳を行い，T の対訳語彙の入力側の単語 w_X それぞれに対して，対応する単語 w_Y が翻訳結果に出力されていれば，その対訳単語対 (w_X, w_Y) が獲得できているとみなす．

翻訳モデルごとに対訳語彙の獲得を比較するために，テストデータの対訳語彙 T に含まれる入力単語延べ数に対する対訳語を獲得できた割合を対訳語彙獲得率と定義し，本研究の評価指標とした．

4　実験

IBT によってドメイン適応を行い，翻訳モデルの翻訳結果からターゲットドメインの対訳語彙獲得を確認することにより，ドメイン知識の獲得を調査する．また，コーパスの前処理や単言語コーパスの違いによって対訳語彙の獲得にどの程度影響を与えるのか調査する．

4.1　データセット

ソースドメインの対訳コーパスには，対訳 440,288 文からなる Wikipedia 日英京都関連文書対訳コーパス (KFTT) を用いた．ターゲットドメインの単言語コーパスには，英語と日本語の対訳コーパスである Asian Scientific Paper Excerpt Corpus (ASPEC)[4] の対訳データの全文 100 万文を 50 万文ずつ分割して，両言語に前半 50 万文を単言語コーパスとして使用したコンパラブルなもの (以降，この単言語コーパス対を **CP** と表記) と，英語の前半 50 万文と日本語の後半 50 万文をそれぞれ使用した非コンパラブルなもの (以降，**NCP** と表記) について調査した．**CP** は各言語の単言語コーパスに，同じコンテキストで対訳語が出現する理想的な条件，**NCP** は必ずしも同じコンテキストで対訳語が出現するとは限らない現実的な条件，に相当する．テストデータと開発データには ASPEC 指定のものを用いた．

コーパスの前処理については，処理単位をサブワードとした場合と単語の場合の 2 通りを実験した．処理単位をサブワードとする場合は，すべてのテキストデータに対して NFKC 正規化を行い，さらに英語データに対しては Moses[3] のトークナイザによる形態素解析と truecaser による小文字化を行った．そして，両言語を Sentence Piece [5] を用いてサブワード単位の分割を行った．ボキャブラリサイズは 16,000 に設定した．処理単位を単語とする場合は，同様に NFKC 正規化を行った後，日本語データには MeCab[6] を用いた形態素解析，英語データに対しては Moses のトークナイザによる形態素解析と truecaser による小文字化を行った．truecaser と Sentence Piece のモデルの学習には，ソースドメインの対訳コーパスとターゲットドメインの単言語コーパスの両方を用いた．

調査対象の対訳語彙を求める際の単語アライメント作成には，ASPEC コーパスに NFKC 正規化，形態素解析，全英単語の小文字化の前処理を行った．英語の形態素解析には Moses のトークナイザを，日本語には MeCab を使用した．その後，Moses を用いた単語アライメントを作成するために，どちらかの言語で 40 単語を超える文を含む文ペアの削除を行った．EM 学習には，ターゲットドメインの利用できる全ての対訳データを用いて単語アライメントを作

1)　単語アライメントの精度を向上させるため，ドメイン適応実験で単言語コーパスとして利用するものも含め、利用できる対訳コーパスを全て用いて EM 学習を行った．

表 1　調査対象対訳語彙の統計

	単言語コーパス対			
	CP		**NCP**	
	En	Jp	En	Jp
D_X の単語タイプ数	1,464	1,074	1,464	1,087
テストデータ中の単語トークン数	806	811	796	805

成し，調査対象のテストデータの単語アライメントだけを実験に利用した．文ペア削除後のテストデータのサイズは 1,557 文対であった．

4.2　実験方法

3 章の手順に従って対訳語彙の特定を行った．まず，KFTT 対訳コーパスに出現する単語集合 O_{En}, O_{Ja}，ASPEC 単言語コーパスに出現する単語集合 I_{En}, I_{Ja}，からターゲットドメインの ASPEC 単言語コーパスにのみ存在する単語 D_{En}，D_{Ja} を求めた．その後，テストデータ間の単語アライメントを作成して，作成したアライメントと D_{En}，D_{Ja} を用いて調査対象の対訳語彙 T を特定した．

翻訳モデルの学習は 2 章で記述した IBT の手順に従った．まず，ソースドメイン対訳コーパス KFTT から翻訳モデル Model0 を学習し，ターゲットドメイン単言語コーパスである ASPEC を Model0 を用いて翻訳して疑似対訳コーパスを作成した．次に，KFTT の対訳コーパスと ASPEC の疑似対訳コーパスを連結した学習データを用いて Model0 を fine-tuning して翻訳モデル Model1 を得た．その後，再び単言語コーパスを Model1 を用いて翻訳して疑似対訳コーパスを作成した．以降もこのような手順を双方向かつ反復的に行うことでモデルの学習を行った．

次に，学習した各種モデルを用いてテストデータを翻訳した．翻訳モデルの処理単位をサブワードとした場合は，翻訳結果のサブワード列をデトークナイズした後，日本語は MeCab による形態素解析を行い，単語列に変換した．得られた単語列に対して，3 章の手順に従って翻訳結果に対訳語彙が出力されているか確認を行った．

4.3　実験条件

ニューラル機械翻訳システムには OpenNMT[7] を用いた．モデルには Transformer を使用し，最適化アルゴリズムには Adam を用いて学習率を 2 とした．学習ステップは Model0 では 30000 ステップ，以降の Model では 5,000 ステップずつ学習を行い続け，開発データに対する accuracy が 3 回連続で向上

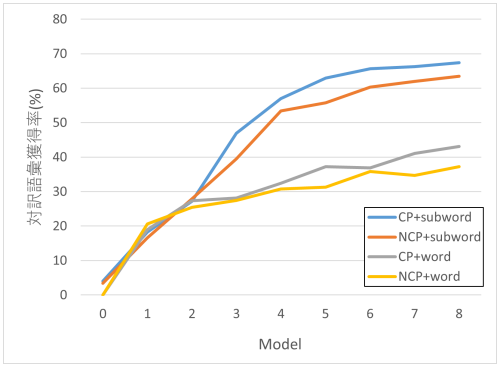

図 2　英日方向の対訳語彙獲得率

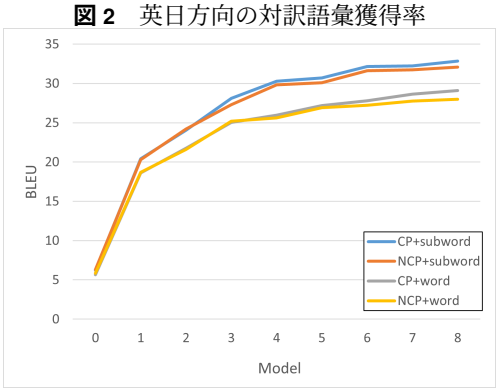

図 3　英日方向の BLEU

しない場合に終了し，もっとも高かった accuracy のモデルを選択した．処理単位を単語とする場合は，ボキャブラリサイズ 50,000 に加えて，調査対象の対訳語彙の単語をすべてボキャブラリに追加した．これは，単語単位の翻訳の場合，そもそもボキャブラリに含まれない単語を出力することはできないため，語彙獲得不可能となってしまうからである．翻訳のボキャブラリには含まれているという条件の下で，対訳語彙獲得を調査した．また，単語単位の翻訳モデルでは，翻訳時に未知語を入力側のもっともらしい単語にそのまま置き換える OpenNMT の replace_unk を使用した．モデルの翻訳精度の評価には BLEU を用いた．

4.4　実験結果

表 1 に調査対象の対訳語彙に関する統計を示す．英日と日英方向の翻訳実験それぞれについて，単言語コーパス対がコンパラブルである場合 (**CP**) と非コンパラブルである場合 (**NCP**)，翻訳単位をサブワードとした場合 (**subword**) と単語とした場合 (**word**)，の組み合わせで $2 \times 2 = 4$ 通りの実験を行った．図 2 と図 3 に，英日方向の対訳語彙獲得率と翻訳性能の結果を，それぞれ示す．また図 4 と図 5 に，日英方向の対訳語彙獲得率と翻訳性能の結果

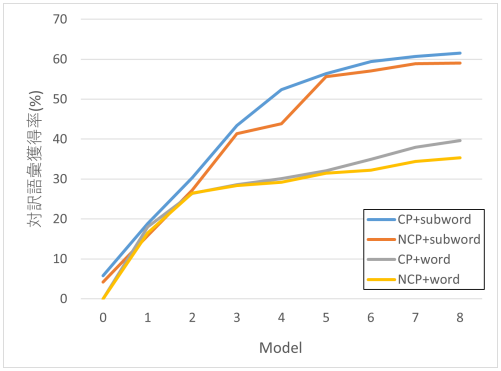

図 4　日英方向の対訳語彙獲得率

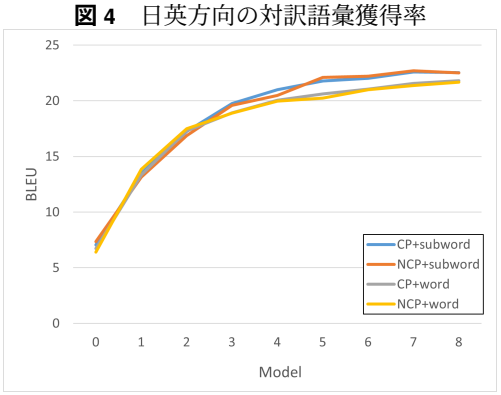

図 5　日英方向の BLEU

を，それぞれ示す．各グラフの横軸は，IBT の反復回数を示し，Model0 がソースドメイン対訳コーパスだけで学習した場合，Model1 が一度だけターゲットドメインの単言語コーパスを逆翻訳した場合 [8]，Model2 以降が IBT に相当する．

実験結果から，どの翻訳方向と実験状況の場合でも，IBT の反復を繰り返すことで対訳語彙獲得率が徐々に上昇するとともに翻訳性能 (BLEU) も向上していることがわかる．理想的な条件である **CP+subword** において，対訳語彙獲得率は英日方向で約 67%，日英方向で 61%となり，獲得可能な 6 割以上の対訳語彙が獲得できた．現実的な条件である **NCP+subword** においても，英日方向で約 63%，日英方向で 59%と 6 割前後の対訳語彙が獲得されていた．この結果から，IBT は対訳語彙の獲得が可能で，その性質が翻訳性能の向上に寄与していると考えられる．また，両言語で対訳語が同じコンテキストで出現するとは限らない現実的な条件である **NCP** でさえ対訳獲得獲得はそれほど低下しなかったことから，IBT は使用する単言語コーパス対に依存せずロバストに語彙獲得できることが示された．

処理単位がサブワードである場合と単語の場合を比較すると，**subword** は **word** よりも高い対訳語彙獲得率を達成した．これにより，対訳語彙獲得には処理単位をサブワードとすることの有効性が示された．実際，Model0，すなわちソースドメインの対訳コーパスのみで学習した場合，ターゲットドメインの対訳は獲得できないはずであるが，それでも **subword** の場合は少量の対訳語彙が獲得できている．これは，サブワードを使用するだけで音訳的な対訳語の獲得 (例えば、albumin とアルブミン、UHV と UHV、など) ができているためである．しかしながら，単語単位であっても IBT の反復によって 4 割程度の対訳語彙獲得が達成できていることから，サブワードの利用が対訳語彙獲得の必要条件ではないことを示している．

表 2　実際に獲得できた対訳語彙の例

英語	日本語
dielectric	誘電
nonlinear	非線形
superconductivity	超電導
broadband	広帯域
ventricular	心室
prognostic	予後
transfusion	輸血
antibody	抗体
histogram	ヒストグラム
biomechanics	バイオメカニクス
MRI	MRI
PCR	PCR

NCP+subword の設定で，実際に獲得できた対訳語彙の例を表 2 に示す．科学技術論文ドメインである ASPEC から，trivial ではない対訳語が多数獲得できていることがわかる．また，IBT の繰り返しによる翻訳結果の変化の例を付録の表 3 に示す．

5　結論

本研究では，IBT のターゲットドメインの知識の獲得過程を，対訳語彙の獲得を確認することで調査を行った．結果として，IBT の反復を繰り返すごとに対訳語彙を獲得していき，最終的にターゲットドメインの獲得可能な対訳語彙の 6 割以上を獲得したことを確認した．同様に BLEU も向上していることから，IBT ではドメイン内コーパスの対訳語彙を獲得することによって，翻訳性能の向上に寄与していることが明らかになった．また，IBT は利用する単言語コーパスがコンパラブルかそうでないかに関わらず，単言語コーパスのみから対訳語彙の単語同士の関係を学習できていること，翻訳の処理単位にサブワードを用いることで対訳語彙獲得率を向上させること，がわかった．

謝辞

本研究は JSPS 科研費 19K11980 および 18H01062 の助成を受けた.

参考文献

[1] Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. Iterative Back-Translation for Neural Machine Translation. In Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pp. 18–24, 2018.

[2] 森田知熙, 秋葉友良, 塚田元. Fine-Tuninng と混成的な逆翻訳サンプリングに基づく NMT の双方向反復的教師なし適応の改善. 言語処理学会 第 27 回年次大会 発表論文集,pp.1669-1673, 2021.

[3] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pp. 177-180, 2007.

[4] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. ASPEC: Asian scientific paper excerpt corpus. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 2204–2208, 2016.

[5] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 66-71, 2018.

[6] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying Conditional Random Fields to Japanese Morphological Analysis, In Proceedings of the Conference on Emprical Methods in Natural Language Processing, pp.230-237, 2004.

[7] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In Proceedings of ACL 2017, System Demonstrations, pp 67-72, 2017.

[8] Rico Sennrich, Barry Haddow, and Alexandra Birch, Improving Neural Machine Translation Models with Monolingual Data, In Proceedings of the Annual Meeting of the Association for Computational Linguistics, pp.86-96, 2016.

A 付録

表 3 対訳語彙獲得の例

ソース	a sacrificial **anode** was successfully developed afterwards , and the safety was raised .
ターゲット（リファレンス）	その後 犠牲 **陽極** の 開発 に 成功 し , 安全 性 を 高め た 。
翻訳結果（Model0）	後 から 犠牲 が 出 て 安全 に なっ た 。
翻訳結果（Model1）	犠牲 的 分子 は その後 成長 し , 安全 性 を 高め た 。
翻訳結果（Model2）	その後 犠牲 **陽極** が 開発 され , 安全 性 が 向上 し た 。
翻訳結果（Model6）	その後 犠牲 **陽極** の 開発 に 成功 し 安全 性 が 高め られ た